IBM CAPSTONE PROJECT

Best location to open an Italian Restaurant in Toronto

1. Description of the problem

The objective of this project is:

To find the most suitable location for an entrepreneur to open an Italian restaurant in Toronto. By using data science methods such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open an Italian restaurant, where should they consider opening it?

2. Description of the data and how it will be used to solve the problem

To solve this problem, the data bellow is need it.

- List of neighborhoods in Toronto.
- Latitude and Longitude of these neighborhoods.
- Venue data related to Italian restaurants.

Extracting the data

- Scrapping of Toronto neighborhoods via Wikipedia.
- Getting latitude and longitude data of these neighborhoods by importing and Excel file.
- Using Foursquare API to get venue data related to these neighborhoods.

3. Methodology

First, I had to extract the list of neighborhoods in Toronto from Wikipedia, using beautifulsoup as bellow:

- wiki_url ='https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
- page = requests.get(wiki_url)
- soup = BeautifulSoup(page.text, 'html.parser')

Having the list of neighborhood names and postal codes, I had to add to my table the coordinates, an Excel sheet was provided, so I just had to import the excel sheet to the notebook and merge the 2 tables together.

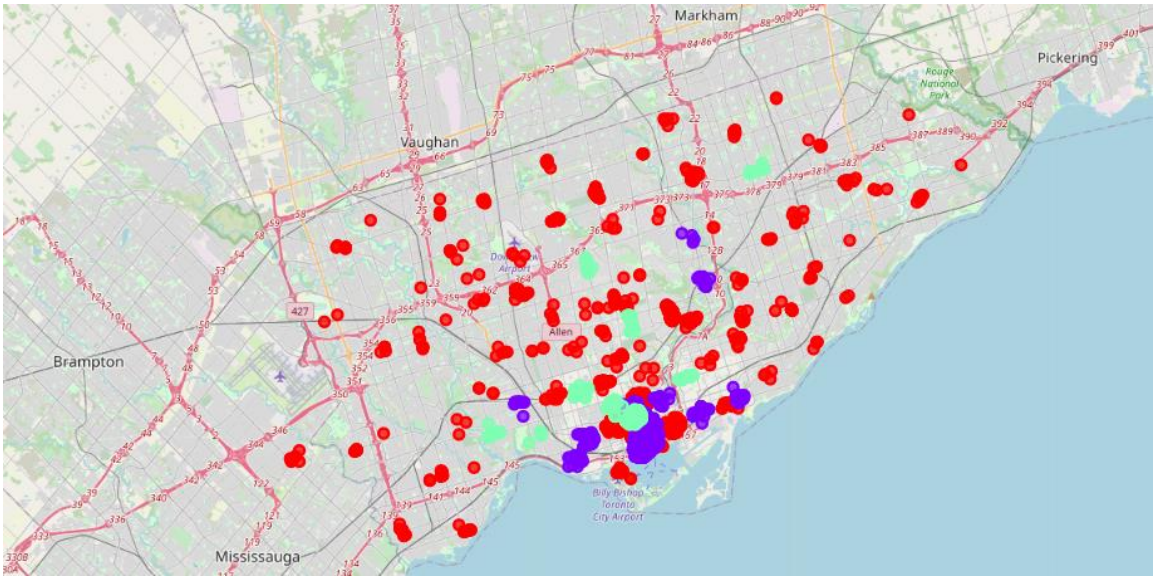
I also did some data cleaning by removing “Not assigned” and “nan” values, and data visualization to check if there are errors with the data.

After having the list of Toronto neighborhood with coordinates I used Foursquare to extract the list of venues in the neighborhoods. Then, I analyzed each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category.

4. Result

The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Italian restaurants are in each neighborhood:

- Cluster 0 (red): Neighborhoods with no Italian restaurants
- Cluster 1 (blue): Neighborhoods with little Italian restaurants
- Cluster 2(green): Neighborhoods with high number of Italian restaurants



5. Observation and Recommendation

Most of Italian restaurants are in Cluster 2 which is around University of Toronto and Little Italy and lowest (close to zero) in Cluster 0 areas which are East York, North York, mainly outside of Toronto downtown. It seems Cluster 1 might be a good location as there are not a lot of Italian restaurants in these areas.

Limitations and Suggestions for Future Research

In this project, I only take into consideration of one factor: the occurrence of Italian restaurants in each neighborhood. There are many other factors that can be taken into consideration such as population density, income of residents, rent, etc. However, for the purpose of the capstone project it's a good start to put into practice what we have learned, and we will need a lot of time to do future research.

6. Conclusion

In this project, we went through the data science process, by identifying the business problem, preparing and analyzing the data, to performing machine learning with K-means clustering, with the purpose to help in decision making, in our project best location to open an Italian restaurants in Toronto.